结合频域信息与对抗网络的虚假图像检测

黄珊珊[1,2], 金 鑫[2,3], 吴 楠[2,3], 江 倩[2,3]

[1]重庆大学大数据与软件学院 重庆 中国 401331
[2]跨境网络空间安全教育部工程研究中心 昆明 中国 650504
[3]云南大学软件学院 昆明 中国 650504

摘要 随着深度学习方法的发展, 深度造假(Deepfake)技术越发成熟。大量近似真实自然的图像涌入人们的生活, 在满足个人娱乐兴趣的同时, Deepfake 技术的滥用对个人隐私、经济市场乃至国家安全构成了潜在威胁。因此, 针对虚假图像的检测方法亟待研究。现有的虚假图像检测技术大多存在准确率低、泛化性差、鲁棒性不足的问题, 因此, 本文从 Deepfake 技术的图像生成机制出发, 对生成的虚假图像存在缺陷进行分析, 并提出了一种基于生成对抗网络的虚假图像检测模型。该模型利用离散傅里叶变换方法将图像从图像域转换到频域, 并将 U-Net 结构和谱归一化引入鉴别器; 利用生成对抗网络优异的特征学习和提取能力, 实现了虚假图像的模式分类。此外, 一种新颖的复合损失函数被提出, 以增强模型检测性能。提出的方法分别在 7 个单独数据集和 1 个混合数据集上进行实验验证, 并采用 3 种实验指标进行模型性能分析。本文方法在单独数据集上最高可达到 100% 准确率, 最低准确率也可达 88.53%; 模型检测召回率, 精确率和 F1 分数平均分别可达 98.17%, 98.25%, 98.19%。此外, 无论是在混合数据集, 还是在模型未知的跨数据集上, 提出方法都能获得良好的模型检测性能。即使在图像压缩的情况下, 本文方法仍然具有较强的鲁棒性。实验与理论结果表明, 与现有先进的虚假图像检测方法相比, 本文方法是一种有效且具有良好泛化性和鲁棒性的虚假图像检测方法。

关键词 信息安全; 图像处理; 深度造假; 虚假图像检测; 生成对抗网络
中图法分类号 TP391.41 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2023.11.04

Fake Image Detection Combining Frequency Domain Information and Adversarial Network

HUANG Shanshan[1,2], JIN Xin[2,3], WU Nan[2,3], JIANG Qian[2,3]

[1] School of Big Data & Software Engineering, Chongqing University, Chongqing 401331, China
[2] Engineering Research Center of Cyberspace, Kunming 650504, China
[3] School of Software, Yunnan University, Kunming 650504, China

Abstract With the development of deep learning methods, deepfake technique has become more and more mature. With the flood of plausible images poured into people's insight, as well as satisfying personal entertainment interests, the misuse of deepfake technique poses a potential threat to personal privacy, economic markets, and even national security. Therefore, it is urgent to study fake image detection methods. Most of the existing methods have the problem of low accuracy, poor generalization, and lack of robustness. Therefore, this paper analyzes the defects of generated false images from the image generation mechanism of deepfake technique, and proposes a fake image detection model based on generative adversarial networks. This model uses the discrete Fourier transform method to transform the image from the image domain to the frequency domain, and introduces the U-Net structure and spectral normalization into the discriminator; the excellent feature learning and extraction capability of generative adversarial networks are utilized to achieve pattern classification of fake images. In addition, a novel composite loss function is proposed to enhance the model detection performance. The proposed method is validated on seven individual datasets and one hybrid dataset, respectively, and three experimental metrics are used for model performance analysis. Compared with the existing advanced methods, the proposed method can achieve the highest accuracy of 100% and the lowest accuracy of 88.53% on separate datasets. The average recall, precision, and F1 score are 98.17%, 98.25%, and 98.19%, respectively. In addition, the proposed method achieves good model detection performance regardless of the hybrid dataset or the cross-dataset where the model is unknown. Further, even in the case of image compression, the method in this paper still has strong robustness. In summary, compared with existing

通讯作者: 江倩, 博士, 副教授, jiangqian_1221@163.com。

本课题得到国家自然科学基金(No. 62002313, No. 62101481, No. 62261060); 云南省重点研发领域计划项目(No. 202001BB050076); 云南省基础研究计划项目(No. 202201AU070033, No. 202201AT070112, No. 202301AU070210, No. 202301AW070007); 云南省科技厅重大科技专项(No. 202202AD080002); 云南省软件工程重点实验室开放基金资助项目(No. 2020SE408); 云南大学第 11 届研究生科研创新项目(No. 2019164)资助。
收稿日期: 2022-03-17; 修改日期: 2022-05-27; 定稿日期: 2023-09-01

state-of-the-art fake image detection methods, the proposed method is an effective fake image detection method with good generalization and robustness.

Key words cyber security; deepfake; fake image detection; generative adversarial networks

1 引言

随着图像处理与深度学习技术的发展, 生成图像的质量大幅提升且获取成本降低, 对社会生活和生产活动产生的威胁日益显著。图像造假方法并非是一项新兴的技术, 最早可以追溯到 19 世纪中期。1860 年, 参议员 John Calhoun 的头部经过人为处理被替换为美国总统 Lincoln 的头部, 从而获得了历史上第一幅假图像。2004 年, George W. Bush 竞选宣传图也是通过人为图像处理获得的假图像, 在选举时干扰民众决策, 影响选举结果。上述事例表明, 图像造假技术的滥用, 对国家的政治生活、宗教安全等可能产生一定程度的影响[1]。近年来, 随着多媒体采集工具的普及, 特别是 photoshop、美图秀秀等图像编辑软件的推广, 数字图像的制作和传播进入了大爆发时代。

特别的, 生成对抗网络 (generative adversarial networks, GANs) 凭借其强大的图像生成能力, 在计算机视觉和图像处理领域[2-7]展现出巨大的应用潜力。也正因为生成对抗网络技术的发展, 大量近似真实、人眼难以辨别真伪的虚假图像开始涌现在大众的视野。2017 年, 首个 deepfake 视频出现, 视频中一位名人的脸被换成了一位色情演员的脸, 对当事人的名誉造成恶劣的影响。2019 年, 一款名为“ZAO”的视频编辑软件出现并受到用户热烈追捧, 该软件可以通过上传自拍照片, 将自己“变身”成为影视片段的主角, 满足了用户猎奇心理的同时, 也带来了一些数据隐私泄露的隐患。不仅如此, Deepfake 技术也可能用于制造虚假经济新闻, 从而造成金融市场混乱; 编辑地球的卫星图像, 以迷惑军事分析人员等等。所有这些形式的造假都有可能对个人隐私乃至社会安定产生巨大的负面影响, 因此迫切需要一种虚假图像检测技术来检测包括卷积神经网络 (convolutional neural networks, CNNs), 特别是 GANs 生成的虚假图像, 该项技术的研究不仅有利于保护个人隐私, 维护社会和谐, 而且具有长远的应用前景和巨大的研究价值。

现有的虚假图像检测方法大致可以分为两类, 一类是利用图像中特定的线索, 比如像素级别的不一致性来进行假图像检测[8]; 而另一类则是基于深度学习模型来捕捉图像中的虚假特征, 从而提高检测性能[9]。特别的, 得益于深度卷积神经网络 (Deep CNNs, DCNNs) 强大的特征提取和表示学习能力, 其在图像分类领域大展身手。虚假图像检测作为二分类问题, 更是使得基于 DNN 的虚假图像检测方法[10-18]成为该领域的主流方法。2018 年, 一种基于自动编码器 (autoencoder, AE) 的假人脸检测模型被 Cozzolino 等人[11]提出, 该模型仅需要少量目标域训练样本就可获得良好的人脸图像检测准确率。尽管如此, 该方法在跨模型获得的不同虚假图像数据集上的泛化性能不足。为了进一步提高虚假人脸图像检测模型的性能, 研究人员开始对视频中的生物学信号进行研究[19-20]。例如, Qi 等人[20]利用生物信号信息来辅助虚假视频检测, 利用光学体积描记术 (photoplethysmography, PPG) 技术, 监测皮肤颜色的极小周期性变化。这是因为真实面部血液会随心跳波动, 而通过生成模型产生的虚假图像会破坏正常的心跳节奏。基于这一推测, 设计了一种用于虚假视频检测的双时空注意网络。

现有的虚假图像检测技术大多是针对人脸图像, 也有一些方法开始专注于 GAN 生成的虚假自然图像检测。例如, Nataraj 等人[13]提出的一种基于 CNN 的虚假图像检测模型, 该模型在像素域提取红、绿、蓝三通道的共现矩阵来进行虚假图像检测。尽管其具有良好的检测准确率, 但由于其只能检测由 CycleGAN 和 StarGAN 模型生成的图像, 从而限制了该模型的应用。针对这一问题, Wang 等人[12]提出了一种通用的虚假图像检测器, 并构建了一种由 11 种不同的基于 CNNs 的图像生成模型生成的虚假图像数据集。此外, Zhang 等人[17]发现常用的上采样方法导致 GANs 模型产生的图像存在独特的伪影, 这种伪影表现为频谱在频域的复制。基于这一发现提出了一种基于谱输入的分类器, 实验表明该分类器在一定程度上提高了虚假图像检测的准确率, 这一发现在文献[18]中得到了进一步的证实。因此许多工作开始专注于基于频域信息的虚假图像检测方法的研究, 例如 Frank 等人[21]和 Dzanic 等人 [22]分别采用不同的频谱变换方法, 包括离散余弦变换(Discrete Cosine Transform, DCT)和离散傅里叶变换(Discrete Fourier Transform, DFT)进行虚假图像检测, 并取得了良好的检测准确率。

因此, 为了充分利用光谱变换和 GAN 模型的优

势, 本文提出了一种基于 GANs 和频谱归一化的虚假图像检测方法。在训练阶段, 该模型首先利用 DFT 将深度网络生成的图像从图像域转换到频域; 并将处理后的结果图像输入到 GANs 的生成器中提取潜在编码和图像重构; 然后将 U-Net 结构引入到鉴别器中, 其中 U-Net 结构包括编码器和解码器两部分, 编码器用于图像分类, 解码器用于对图像像素进行判别, 从而增强鉴别器的鉴别能力。在测试阶段, 利用训练好的鉴别器进行虚假图像的检测。另外本文设计了一种复合损失函数以更好的进行模型优化, 为了全面评估本文方法的有效性, 本文采用多种评价指标, 实验表明本文方法能够取得良好的虚假图像检测性能。

2 基于生成对抗网络的虚假图像检测方法

本节介绍了提出的基于 GANs 的虚假图像检测模型, 该模型主要由三部分构成: 频域转换模块, 生成器模块和鉴别器模块, 总体结构如图 1 所示。另外, 本文除了采用对抗损失来优化生成器和鉴别器模型, 还采用了标签损失来约束生成潜在编码与图像标签间的一致性; 同时, 为了保证重构图像与输入图像的相似性, 本文通过采用重构损失来避免重构图像过程中的像素级损失。下面将对模型的三大模块以及提出的损失函数进行详细阐述。

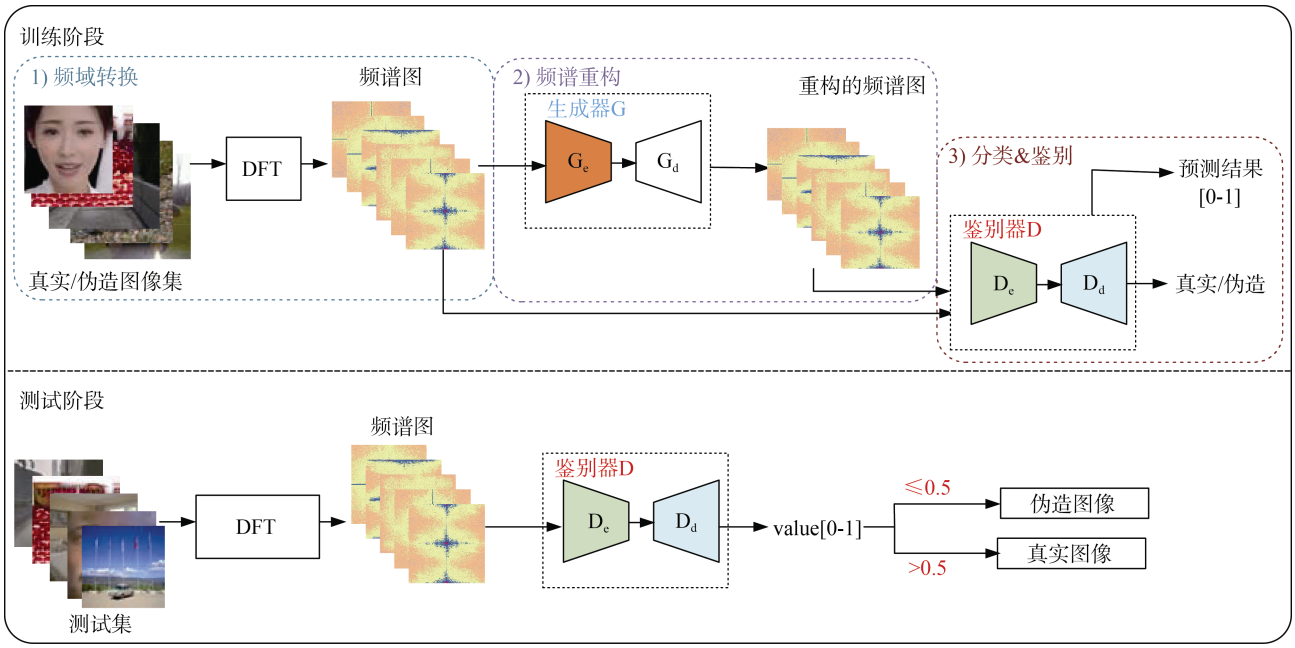

图 1 总体架构

Figure 1 Overall structure of the proposed method

2.1 频域转换

图 2 中展示了由不同生成模型生成的虚假图像对应频谱图, 第一行和第二行分别表示真实图像和由不同生成模型产生的虚假图像。提出方法首先对真实图像和基于不同生成模型获得的虚假图像进行 DFT, 然后对获得的频谱图进行中心化, 接着将中心化后的频谱图输入到提出模型中。其中傅里叶变换如公式(1)所示。

$$F(l,k)=\frac{1}{MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}f(m,n)e^{-j2\Pi\left(\frac{lm}{M}+\frac{kn}{N}\right)} \quad (1)$$

其中, $l=m=0,1,2,\cdots,M-1$, $k=n=0,1,2,\cdots,N-1$。$f(m,n)$ 表示图像信号函数, $\boldsymbol{M}\times\boldsymbol{N}$ 表示图像的尺寸。在本文中, $\boldsymbol{M}=\boldsymbol{N}=256$。

之所以将输入图像从图像域转换到频域, 最重要的一点在于, 通过深度学习模型生成的虚假图像, 从图像域的角度来看往往难以发现造假的痕迹, 特别是针对于近年来提出的例如 StyleGAN、StyleGAN2 等模型, 其往往能够生成高质量、高分辨率图像; 而将图像从图像域转换到频域, 可以发现, 真实图像和由生成模型获得的虚假图像存在很大差异。特别的, 观察图 2 可以发现, 真实图像的高频信息明显要高于虚假图像中的高频信息量。这是因为对于生成模型而言, 伪造图像中灰度变化缓慢的区域往往能够很好生成, 而对于灰度信息变化剧烈的区域, 即边缘区域, 往往很难生成。

2.2 网络结构

本小节对提出模型中的生成器结构和鉴别器结构进行了详细介绍, 并在第 3 章对提出模型结构的有效性进行了实验验证。

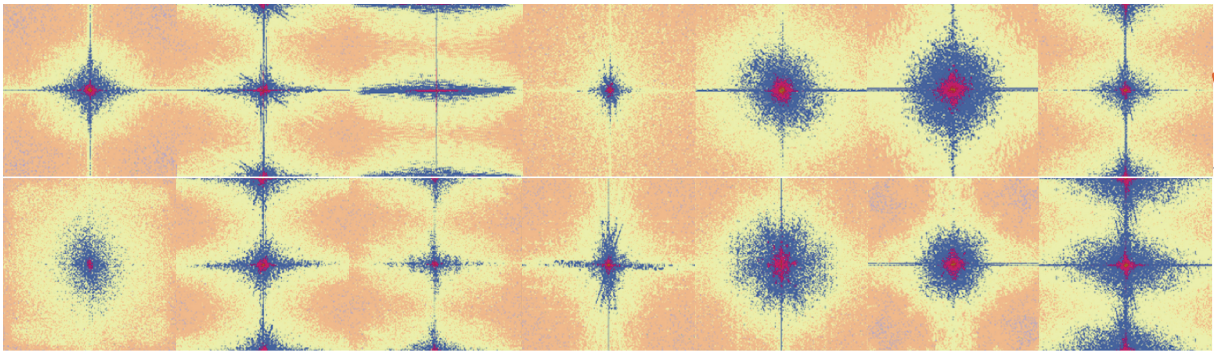

图 2　真实图像和不同模型生成的虚假图像对应频谱图

Figure 2　Corresponding spectrum of real images and fake images generated by different models

图 3 中展示了提出模型中生成器的结构, 该生成器由编码器 (即图 3 左侧的网络) 和解码器 (即图 3 右侧网络)两部分构成, 其中编码器部分用于提取输入图像的潜在编码, 解码器用于图像重构。为了使得重构图像与输入图像尽可能的相似, 本文采用了标签损失来约束潜在编码和标签的一致性, 同时采用重构损失来约束重构图像与输入图像的相似度。损失函数将在第 2.3 节进行详细介绍。生成器模型中编码器由 5 个卷积模块构成, 其中前四个卷积模块由一个卷积层 (convolution, Conv) 和一个批处理层 (Batch normalization, BN) 后跟一个 Leaky-Rectified Linear Units (LReLU) 激活层构成, 最后一个卷积模块则将 LReLU 激活层替换为 Rectified Linear Units (ReLU) 激活层。解码器结构与编码器结构相对应, 由 5 个反卷积模块构成, 每个反卷积模块都具有相同的结构, 包括一个反卷积层和一个 ReLU 激活层。

受到 Schönfeld 等人[23]研究工作的启发, 提出模型中鉴别器的结构与大多数现有的鉴别器都不同, 其采用了 U-Net 结构, 包括编码器和解码器两部分, 其中编码器用于对输入图像进行分类, 即将输入图像分类为真实图像或者虚假图像; 解码器用于对图像进行逐像素判别, 将图像分割为真实区域和虚假区域, 从而使得鉴别器能够同时学习到真实图像和生成的虚假图像之间的全局差异和局部差异。因此, 该鉴别器具有两个输出, 对应于图像检测结果和图像像素判别结果。在模型训练阶段, 鉴别器与生成器一起训练; 在测试阶段则仅使用鉴别器进行图像真伪检测。鉴别器模型结构如图 4 所示, 其中编码器和解码器中的具体构成也展示在图 4 中。

特别的, 本文将谱归一化[24] (Spectral Normalization, SN) 应用到鉴别器中, 以一种更优雅的方式使得鉴别器 D 满足利普希茨连续性, 限制了函数变化的剧烈程度, 从而使模型更稳定。在实现 SN 之前, 首先需要求解卷积网络中的卷积核 (权重矩阵) W的奇异值, 从而获得每层参数矩阵的谱范数, 在这一过程中, 本文延续了文献[24]中的做法, 采用“幂迭代法”来近似求取, 其迭代过程如行步骤 2.2 和 2.3 所示; 在求得谱范数之后, 每个参数矩阵上的参数除以它, 以达到归一化目的。算法具体流程如下所示:

算法 1: 谱归一化

1. 利用随机向量, 初始化 $\tilde{u}_l \in \mathcal{R}^{d_l}$ for $l=1,\cdots,L$

2. 对于每一次更新和每一层 l:

2.1 对非归一化权值 W^l 应用幂迭代法:

$$\tilde{v}_l \leftarrow \left(W^l\right)^{\mathrm{T}} \tilde{u}_l / \left\|\left(W^l\right)^{\mathrm{T}} \tilde{u}_l\right\|_2$$

$$\tilde{u}_l \leftarrow W^l \tilde{v}_l / \left\|W^l \tilde{v}_l\right\|_2$$

2.2 通过谱范数 $\sigma\left(W^l\right)$ 计算 $\bar{W}_{\mathrm{SN}}$:

$$\bar{W}_{\mathrm{SN}}^l\left(W^l\right)=W^l/\sigma\left(W^l\right), \text{ where } \sigma\left(W^l\right)=\tilde{u}_l^{\mathrm{T}} W^l \tilde{v}_l$$

2.3 在数据集 S 上利用 SGD 更新 W^l, 其中学习率为 α: $W^l \leftarrow W^l - \alpha \nabla_{W^l} \ell\left(\bar{W}_{\mathrm{SN}}^l\left(W^l\right), \mathrm{S}\right)$

2.3　损失函数

为了加快模型优化速度, 提高模型检测虚假图像的性能, 本文设计了一种复合损失函数, 该损失函数由三个部分构成: 对抗损失, 标签损失和重构损失。

1) 对抗损失

由于本文提出模型是一种基于 GANs 的检测模型, 因此在提出方法中继续沿用了原始 GANs 中的交叉熵损失, 该损失函数能够最大化鉴别器 D 的区分度, 最小化生成器 G 的输出和真实数据的区别。对抗损失的计算公式如(2)(3)所示。

$$\mathcal{L}_{\mathrm{D}} = -\mathbb{E}_x[\log \mathrm{D}(x)] - \mathbb{E}_z[\log(1-\mathrm{D}(\mathrm{G}(z)))] \quad (2)$$

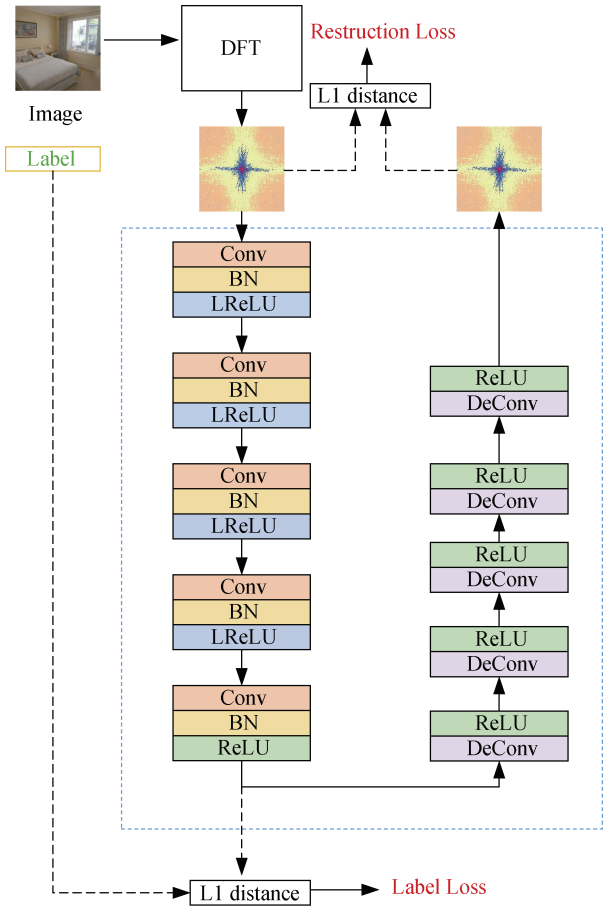

图 3 生成器结构

Figure 3 Structure of the Generator

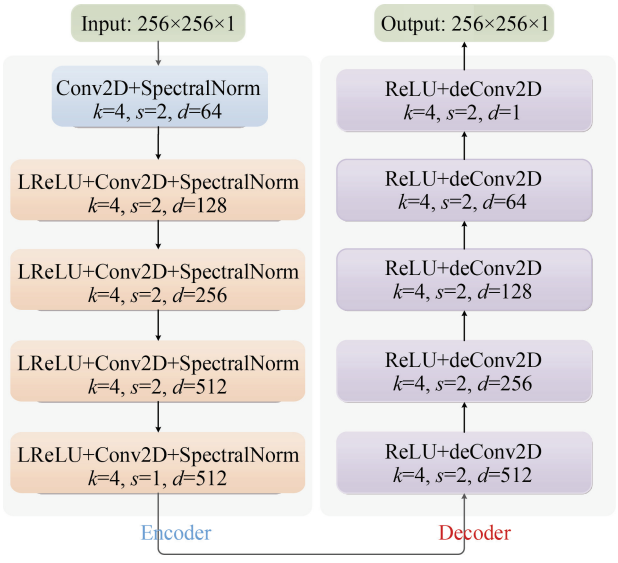

图 4 鉴别器结构

Figure 4 Structure of the discriminator

$$\mathcal{L}_{\mathrm{G}} = -\mathbb{E}_z[\log \mathrm{D}(\mathrm{G}(z))] \tag{3}$$

更近一步的，提出的方法中将 U-Net 结构引入了鉴别器，因此鉴别器的损失通过 $\mathrm{D_d}$ 和 $\mathrm{D_e}$ 的决策共同计算获得，其计算公式如下所示。

$$\mathcal{L}_{\mathrm{D_e}} = -\mathbb{E}_x\left[\log \mathrm{D_e}(x)\right] - \mathbb{E}_z\left[\log\left(1-\mathrm{D_e}(\mathrm{G}(z))\right)\right] \tag{4}$$

$$\mathcal{L}_{\mathrm{D_d}} = -\mathbb{E}_x\left[\sum_{i,j}\log\left[\mathrm{D_d}(x)\right]_{i,j}\right] - \mathbb{E}_z\left[\sum_{i,j}\log\left(1-\left[\mathrm{D_d}(\mathrm{G}(z))\right]_{i,j}\right)\right] \tag{5}$$

其中，$\left[\mathrm{D_d}(x)\right]_{i,j}$ 和 $\left[\mathrm{D_d}(\mathrm{G}(z))\right]_{i,j}$ 分别表示鉴别器在像素 (i,j) 处的决策。对应的生成器损失为:

$$\mathcal{L}_{\mathrm{G}} = -\mathbb{E}_z\left[\log \mathrm{D_e}(\mathrm{G}(z)) + \sum_{i,j}\log\left[\mathrm{D_d}(\mathrm{G}(z))\right]_{i,j}\right] \tag{6}$$

2) 标签损失

标签损失是为了约束通过生成器中编码器 $\mathrm{G_e}$ 编码获得的潜在编码与图像真实标签之间的差异，从而使得重构图像能够与输入图像更加一致，其计算公式如式(7)所示。另外还计算了鉴别器中编码器 $\mathrm{D_d}$ 的输出与标签之间的交叉熵损失，从而辅助鉴别器更好鉴别输入图像的真伪，如公式(8)所示。

$$\mathcal{L}_{\mathrm{label\_G}} = \left|l - \mathrm{G_e}(x)\right| \tag{7}$$

$$\mathcal{L}_{\mathrm{label\_D}} = -\left[l\log\left(\mathrm{D_e}(x)\right) + (1-l)\log\left(1-\mathrm{D_e}(x)\right)\right] \tag{8}$$

其中，l 表示输入图像的标签，x 表示输入图像。

3) 重构损失

此外，为了使得生成器生成图像能够高度还原输入图像，本文还提出了一种重构损失，该损失函数计算了重构图像 $\mathrm{G}(x)$ 和输入图像 x 之间的像素级损失，用 L1 损失表示，可以通过公式(9)计算获得。

$$\mathcal{L}_{\mathrm{recon}} = \frac{1}{MN}\sum_{i=0}^{M}\sum_{j=0}^{N}\left|x_{(i,j)} - \mathrm{G}(x)_{(i,j)}\right| \tag{9}$$

其中，M，N 表示图像的尺寸。

总的损失函数为上述损失函数的加权，表示为:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{G}} + \mathcal{L}_{\mathrm{D_e}} + \mathcal{L}_{\mathrm{D_d}} + \lambda_{\mathrm{label}}\left(\mathcal{L}_{\mathrm{label\_G}} + \mathcal{L}_{\mathrm{label\_D}}\right) + \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{recon}} \tag{10}$$

其中，λ_{label}、λ_{rec} 分别表示标签损失和重构损失的权重。

3 实验结果与分析

为了验证提出方法的性能，本文在不同数据集上开展了实验验证，并通过一系列消融实验证明了提出方法的有效性；另外，本部分还与现有的部分先进的虚假图像检测模型进行了对比实验。本小节首先在第 3.1 小节介绍了模型参数设置和数据集；紧接着在第 3.2 节阐述了模型的评价标准；然后在第 3.3 节对提出方法在不同数据集上的检测性能进行了详细分析，同时展示了与其他现存先进方法的检测

性能对比结果。

3.1 数据集与模型参数

本文使用了 Wang 等人[12]提供的虚假图像数据集进行模型训练和验证, 如图 5 所示。该数据集包含 11 种通过不同的基于 CNNs 的生成模型获得的虚假图像集。本文中仅采用了 8 种虚假图像数据集进行模型验证, 包括 StyleGAN[25], BigGAN[26], CycleGAN[27], StarGAN[28], CRN[29], IMLE[30], StyleGAN2[31], 和 GauGAN[32]。其中前 7 种数据集不仅用于验证检测方法在单独数据集上的检测性能, 而且也应用于在混合数据集上的性能验证。每个数据集的真实图像和虚假图像的比例为 1∶1。在选定的数据集中, 80%的数据用于训练, 其余 20%用于验证所提模型的有效性。不同数据集的详细信息如表 1 所示。实验中关键参数设置如下: 用于训练的图像尺寸为 256×256, 学习率为 0.0002, 训练次数为 100epoch, batchsize 为 1。

3.2 评价指标

为了客观全面的评价提出检测模型的有效性, 本文采用了 4 种常用的模型检测性能评估指标, 包括准确率 (accuracy, ***ACC***), 精确率 (Precision, ***P***) , 召回率 (Recall, ***R***) 和 ***F1*** 分数 (F1-Measure, ***F1***) 。

ACC: ***ACC*** 表示虚假图像检测准确率。对于给定的测试数据集, 分类器正确分类的样本数与样本总数的比值即为准确率。一般情况下, 准确率的值越大, 表示检测模型的性能越优, 其计算公式为:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

其中, *TP* 表示生成图像被判定为假图像, *TN* 表示真实图像被判定为真实图像的情况; *FP* 表示真实图

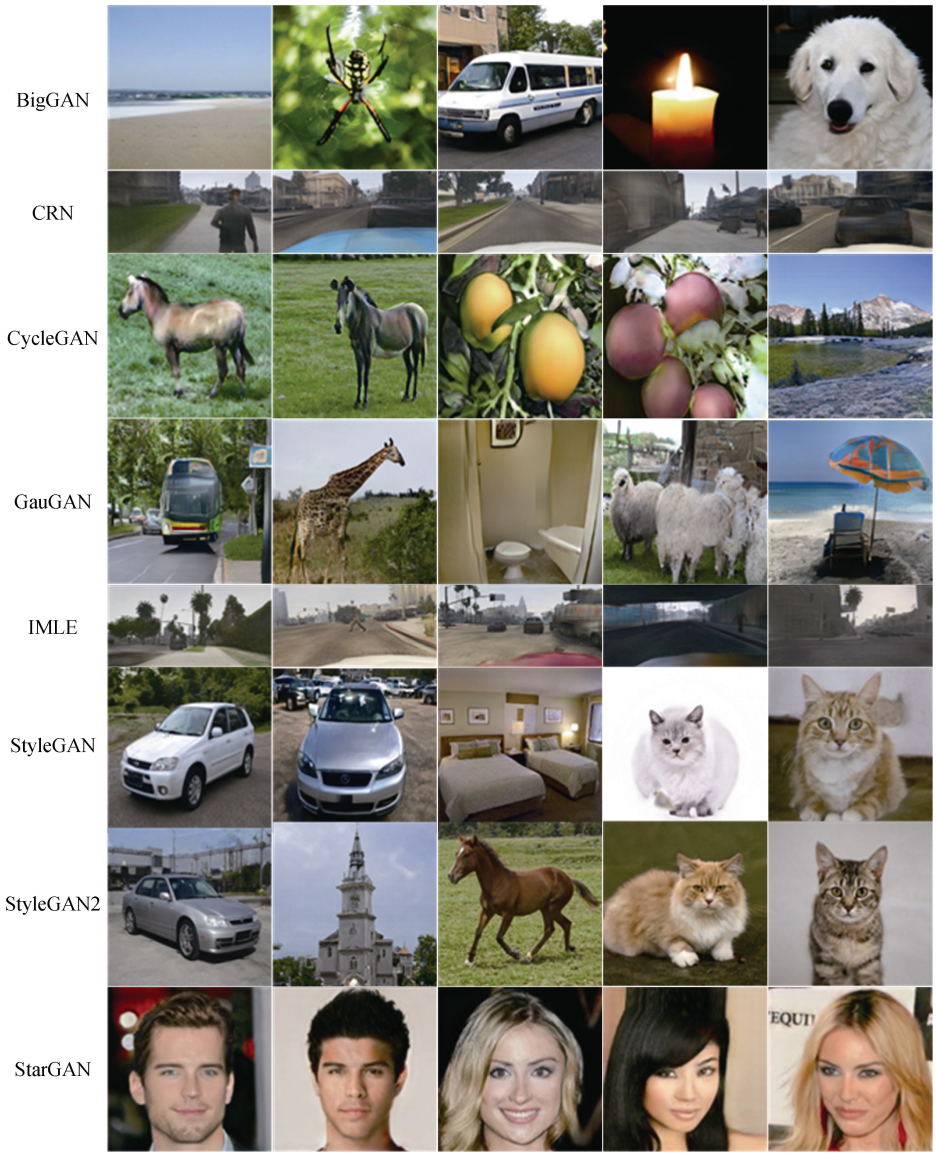

图 5 数据集展示

Figure 5 Presentation of the dataset

表 1 不同虚假图像数据集中数据量

Table 1 Data volume in different fake image datasets

数据集	BigGAN	CRN	CycleGAN	StarGAN	StyleGAN	StyleGAN2	IMLE	GauGAN
Total	4 000	12 764	2 690	3 998	11 982	15 484	12 764	10 000
Train sets	3 000	9 000	1 946	3 000	8 982	11 976	9 000	7 000
Test sets	1 000	3 764	744	998	3 000	3 508	3 764	3 000

像被判定为虚假图像, FN 表示虚假图像被判定为真实图像的情况。

P: ***P*** 为精确率, 又称查准率, 即正确预测为假图像的数量占全部预测为假图像的比例。***P*** 可以通过公式 (12) 计算获得。

$$\boldsymbol{P}=\frac{TP}{TP+FP} \tag{12}$$

R: ***R*** 为召回率, 又称查全率, 即在实际为假图像的所有样本中, 被预测为假图像的概率。***R*** 值越高, 表示实际的假图像被预测正确的概率越高, 可以通过如下公式计算获得。

$$\boldsymbol{R}=\frac{TP}{TP+FN} \tag{13}$$

F1: ***F1*** 分数为精确率和召回率的调和均值, 可以在准和全两方面找到较为综合的效果其值越大越好, 其计算公式可以表示为:

$$\boldsymbol{F1}=\frac{2TP}{2TP+FP+FN}=\frac{2\boldsymbol{PR}}{\boldsymbol{P}+\boldsymbol{R}} \tag{14}$$

3.3 实验结果

为了充分的验证本文提出方法的有效性, 将提出模型在不同的虚假图像数据集上, 针对是否具有频域转换模块、网络结构中是否采用频谱归一化操作以及不同鉴别器的结构进行了一系列消融实验。同时还和现有的部分检测模型进行了对比实验, 并采用了多种不同的客观评估指标对不同模型获得的检测结果进行了对比分析。实验表明, 本文提出模型能够较为准确的检测出虚假图像, 并且具有一定的泛化性。

1) 频域转换的实验验证

表 2～表 5 中展示了提出方法中有无 DFT 和 SN 操作的实验结果对比, 观察表中数据可以发现, 添加频域转换及 SN 操作之后, 模型检测准确率 ***ACC***、精确率 ***P***、召回率 ***R*** 以及 ***F1*** 分数得到普遍提升, 特别是对于 StyleGAN、BigGAN、StarGAN 这三种模型, 未进行 DFT 转换和 SN 操作时, 其准确率 ***ACC*** 和精确率 ***P*** 极低, 表明此时检测模型对于这几种模型产生的虚假图像几乎没有任何作用, 然而在添加频域转换和 SN 操作后, 其检测准确率 ***ACC*** 和精确率 ***P*** 得到大幅提升, 同时召回率 ***R*** 也保持在较优的水平。本文最终方法在对单个数据集进行检测时, 模型召回率 ***R***, 精确率 ***P*** 和 ***F1*** 分数平均分别可达 98.17%, 98.25%, 98.19%, 实验表明, 在频域中, 通过生成模型获得的虚假图像中存在的缺陷更容易被检测。

表 2 有无 DFT 和 SN 操作的模型准确率 (*ACC*) 对比

Table 2 Comparison of model accuracy (*ACC*) with or without DFT and SN operations

评价指标	有无 DFT 和 SN	IMLE	StyleGAN2	StyleGAN	BigGAN	CRN	StarGAN	CycleGAN
ACC	无	**0.998 4**	0.913 3	0.500 0	0.500 0	**0.999 7**	0.500 0	0.551 1
	有	0.885 3	**0.948 5**	**0.945 7**	**0.984 0**	0.993 1	**1.000 0**	**1.0000**

表 3 有无频域转换和 SN 操作的模型精确率 (*P*) 对比

Table 3 Comparison of model precision (*P*) with or without DFT and SN operations

评价指标	有无 DFT 和 SN	IMLE	StyleGAN2	StyleGAN	BigGAN	CRN	StarGAN	CycleGAN
P	无	**0.997 9**	0.923 9	0.500 0	0.500 0	**0.999 5**	0.5000 0	0.530 9
	有	0.859 0	**0.946 3**	**0.975 1**	**0.989 9**	0.996 8	**1.000 0**	**1.000 0**

表 4 有无 DFT 和 SN 操作的模型召回率 (*R*) 对比

Table 4 Comparison of model recall (*R*) with or without DFT and SN operations

评价指标	有无 DFT 和 SN	IMLE	StyleGAN2	StyleGAN	BigGAN	CRN	StarGAN	CycleGAN
R	无	**0.998 9**	0.917 0	**1.000 0**	**1.000 0**	**0.999 5**	1.000 0	0.876 3
	有	0.922 0	**0.951 0**	0.914 7	0.978 0	0.994 0	**1.000 0**	**1.000 0**

表 5 有无 DFT 和 SN 操作的模型 *F1* 分数对比

Table 5 Comparison of *F1* scores of models with or without DFT and SN operations

评价指标	有无 DFT 和 SN	IMLE	StyleGAN2	StyleGAN	BigGAN	CRN	StarGAN	CycleGAN
F1	无	**0.998 4**	0.920 2	0.666 7	0.666 7	**0.998 9**	0.666 7	0.661 2
	有	0.889 4	**0.938 1**	**0.944 0**	**0.983 8**	0.995 4	**1.000 0**	**1.000 0**

2) 网络结构的实验验证

本部分对比了提出网络结构中鉴别器结构是否采用 U-Net 结构, 这里的 U-Net 结构包括编码器和解码器两部分, 具体模块构成如图 4 所示。实验结果通过图表的形式直观的展示在图 6-图 9 中。观察图表可以发现, U-Net 结构的采用在一定程度上提高了模型检测的性能, 其中模型检测准确率 ***ACC***、精确率 ***P***、召回率 ***R*** 及 ***F1*** 分数分别平均提升 3.698%, 2.039%, 6.064%, 4.241%, 这主要是得益于 U-Net 结构强大的特征提取能力。实验表示, 提出模型中采用的编码器-解码器结构相比于普通鉴别器结构更适用于虚假图像检测。

3) 混合数据集及跨数据集实验结果

另外, 本项工作还在混合数据集上进行了模型训练, 然后在不同数据集上进行了单独测试, 其中混合数据集中共有图像 4000 张, 真实图像和虚假图像比例为 1∶1。该混合数据集共包含由五种不同生成模型产生的虚假图像: IMLE、StyleGAN、StyleGAN2、CRN、GauGAN, 各 8000 张图像。为了充分展示本文提出检测模型的泛化性能, 对实验结果进行了综合分析, 结果展示在表 6 中。实验表明, 本文方法在混合数据集上也能够取得较优的检测性能, 其平均

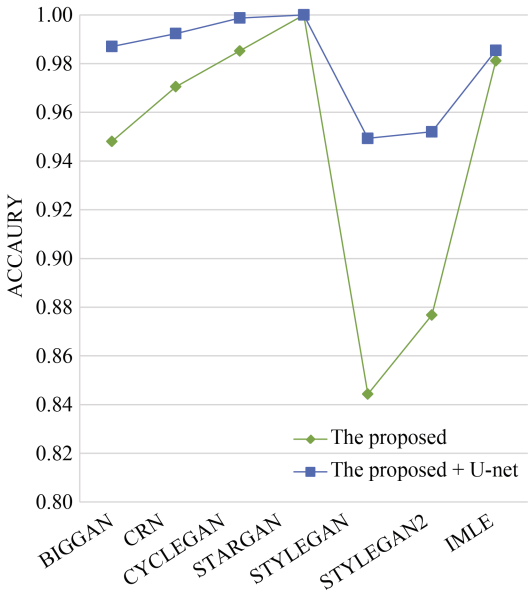

图 6 鉴别器是否采用 U-Net 结构的检测准确率

Figure 6 Detection accuracy of the discriminator with or without U-Net structure

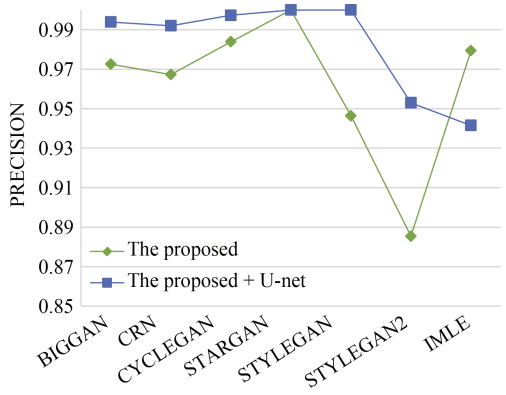

图 7 鉴别器是否采用 U-Net 结构的检测精确率

Figure 7 Precision of the discriminator with or without U-Net structure

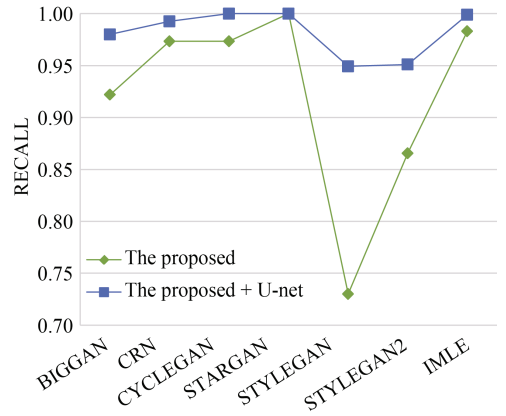

图 8 鉴别器是否采用 U-Net 结构的检测召回率

Figure 8 Recall of the discriminator with or without U-Net structure

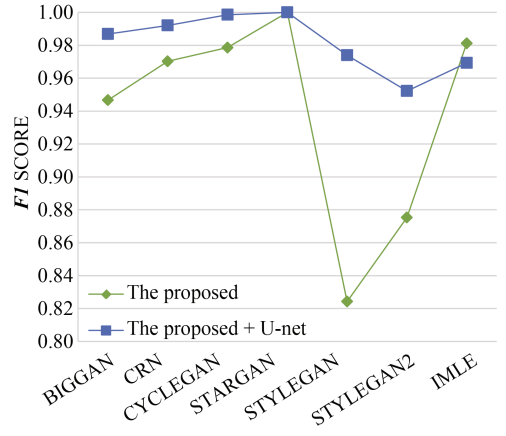

图 9 鉴别器是否采用 U-Net 结构的检测 *F1* 分数

Figure 9 *F1* score of the discriminator with or without U-Net structure

表 6 混合数据集实验结果

Table 6 Experimental results of the hybrid dataset

数据集	*ACC*	*P*	*R*	*F1*
IMLE	0.988 0	0.983 0	0.992 6	0.988 5
StyleGAN2	0.842 3	0.918 7	0.751 0	0.826 4
StyleGAN	0.925 7	0.911 7	0.942 7	0.926 9
CRN	0.979 0	0.965 9	0.993 1	0.979 3
GauGAN	0.913 0	0.896 4	0.934 0	0.914 8
均值	**0.929 6**	**0.935 1**	**0.922 7**	**0.927 2**

检测准确率可达 92.96%, 进而展现出本文提出方法的检测泛化性。

另外, 为了验证本文提出方法在跨数据集上的泛化性, 我们利用基于混合数据集的预训练模型, 来验证本文方法在未知数据集上的性能表现。为此, 本文对由 BigGAN、CycleGAN、StarGAN 生成的虚假图像数据集分别进行了测试; 同时我们利用在 CycleGAN 生成的图像数据集上获得的预训练模型对 StarGAN 生成的图像数据集上进行了测试, 实验结果展示在表 7 中。观察表中数据, 可以发现即使在模型未见过的数据集上, 本文模型仍能够获得良好的检测性能。

表 7 跨数据集实验结果对比

Table 7 Comparison of experimental results cross-datasets

数据集(训练集)	*ACC*	*P*	*R*	*F1*
BigGAN (混合)	0.818 0	0.805 8	0.838 0	0.821 6
StarGAN (混合)	0.810 6	0.910 1	0.689 4	0.784 5
StarGAN (CycleGAN)	0.995 0	0.996 0	0.994 0	0.995 0
CycleGAN (混合)	0.696 0	0.851 9	0.474 2	0.609 3

4) 对比实验

表 8 中展示了本文方法与目前现有部分先进方法的实验对比分析, 其中加粗字体部分表示各项指标的最优值。对比方法包括 Inception[32]、Resnet50[34]、Xception[35] 、Mesonet[36] 、Mesonet-Inception[36] 和 EfficientNet[37]、GANDCT[21]和 FourierSpectrum[22]八种对比算法。表中数据表明, 本文方法在大多数数据集上都能展现出优秀的检测性能, 尽管与基于像素特征的方法相比, 在 CRN、StyleGAN 两个数据集上略低于 Inception 和 EfficientNet 模型, 但也取得了较优的检测准确率; 更进一步的, 与基于频域特征的方法相比, 本文提出的方法仅在 GauGAN 数据集上略低于 GANDCT, 但在其余数据集上仍能取得较高的检测准确率。总得来说, 本文方法具有良好的检测性能。

此外, 为了验证本文方法在图像压缩情况下的模型鲁棒性, 我们分别在三种不同的图像压缩率(50%, 75%, 95%)下进行了实验, 实验结果表明, 随着压缩率的不断增加, 模型的检测准确率会出现逐步的下降。即便如此, 本文方法仍然能取得较好的检测准确率: 在不同图像压缩率(50%, 75%, 95%)下分别可以获得 0.7257, 0.7943, 0.9433 的平均检测准确率。特别的, 在图像压缩率为 75%时, 本文方法与其他基于频域的方法实验对比展示在表 9 中。观察表中数据可以发现, GANDCT 在由 BigGAN, CycleGAN 生成的虚假图像集上能够获得较优的检测准确率, 但在其他数据集上均低于本文提出的方法, 且其他两种基于频域的方法在平均检测准确率上均低于本文方法。总的来说, 即使在图像压缩的情况下, 本文方法仍然具有较强的鲁棒性。

4 结语

本文将频域转换引入了虚假图像检测方法, 同时提出了一种具有 U-Net 结构的鉴别器, 其中编码器用于图像分类, 解码器用于图像像素判别; 该鉴别器还采用了 SN, 使得模型训练更加稳定; 另外设

表 8 不同检测方法的准确率(*ACC*)对比

Table 8 Comparison of accuracy (*ACC*) of different detection methods

数据集	Inception	Resnet50	Xception	Mesonet	Mesonet-Inception	EfficientNet	GANDCT	FourierSpectrum	本文方法
BigGAN	0.500 0	0.500 0	0.500 0	0.526 0	0.739 0	0.696 3	0.939 2	0.796 3	0.984 0
CRN	0.998 8	0.998 0	**1.000 0**	0.958 0	0.9947	0.996 8	0.980 2	0.943 7	0.993 1
CycleGAN	0.652 9	0.589 2	0.546 1	0.661 5	0.627 1	0.628 3	0.987 8	0.986 2	**1.000 0**
GauGAN	0.660 0	0.640 0	0.669 0	0.719 3	0.815 7	0.725 1	**0.924 4**	0.835 9	**0.885 3**
StarGAN	0.890 9	0.500 0	0.998 7	0.531 1	0.999 0	0.974 6	0.862 9	0.962 5	**1.000 0**
StyleGAN	0.495 3	0.500 0	0.519 5	0.693 0	0.797 7	**0.956 8**	0.934 8	0.854 6	0.945 7
StyleGAN2	0.831 8	0.808 5	0.861 5	0.764 3	0.855 0	0.931 5	0.838 8	0.876 9	**0.948 5**
IMLE	0.998 0	0.997 2	0.998 2	0.776 6	0.977 7	0.996 8	0.983 4	0.998 6	**1.000 0**

表 9　不同方法在 75%图像压缩率下的模型鲁棒性对比(*ACC*)

Table 9　Comparison of model robustness of different methods with 75% image compression rate (*ACC*)

对比方法	BigGAN	CRN	CycleGAN	GauGAN	StyleGAN	StyleGAN2	StarGAN	平均值
GANDCT	**0.812 5**	0.845 5	**0.916 7**	0.526 0	0.666 7	0.710 1	0.838 5	0.759 4
FourierSpectrum	0.712 6	0.885 1	0.612 4	0.606 4	0.701 2	0.765 5	0.808 1	0.727 3
提出方法	0.773 0	**0.955 1**	0.646 0	**0.660 1**	**0.784 7**	**0.829 5**	**0.911 8**	**0.794 3**

计了一种复合损失函数进行模型优化, 该损失函数的提出在一定程度上提高了模型检测的准确率。通过一系列的消融实验和对比实验验证了本文提出方法在不同的模型生成的虚假图像数据集上的检测性能, 实验表明本文方法能够达到较高的检测准确率且具有一定程度的泛化性。

由于目前深度造假技术发展迅速, 生成模型生成的图像也愈发逼真, 本文提出方法并不能涵盖所有的通过生成模型产生的虚假图像。因此, 下一步我们将对本文方法进行进一步改进, 研究由生成模型生成的虚假图像存在的通病, 从而探索出一种通用的虚假图像检测模型, 进一步提高检测方法的泛化性能。

参考文献

[1] Li X R, Ji S L, Wu C M, et al. Survey on Deepfakes and Detection Techniques[J]. *Journal of Software*, 2021, 32(2): 496-518.
(李旭嵘, 纪守领, 吴春明, 等. 深度伪造与检测技术综述[J]. *软件学报*, 2021, 32(2): 496-518.)

[2] Chen F J, Zhu F, Wu Q X, et al. A Survey about Image Generation with Generative Adversarial Nets[J]. *Chinese Journal of Computers*, 2021, 44(2): 347-369.
(陈佛计, 朱枫, 吴清潇, 等. 生成对抗网络及其在图像生成中的应用研究综述[J]. *计算机学报*, 2021, 44(2): 347-369.)

[3] Zhai L M, Jia J, Ren W X, et al. Recent Advances in Deep Learning for Image Steganography and Steganalysis[J]. *Journal of Cyber Security*, 2018, 3(6): 2-12.
(翟黎明, 嘉炬, 任魏翔, 等. 深度学习在图像隐写术与隐写分析领域中的研究进展[J]. *信息安全学报*, 2018, 3(6): 2-12.)

[4] Huang S S, Jiang Q, Jin X, et al. Semi-Supervised Remote Sensing Image Fusion Method Combining Siamese Structure with Generative Adversarial Networks[J]. *Journal of Computer-Aided Design & Computer Graphics*, 2021, 33(1): 92-105.
(黄珊珊, 江倩, 金鑫, 等. 结合双胞胎结构与生成对抗网络的半监督遥感图像融合[J]. *计算机辅助设计与图形学学报*, 2021, 33(1): 92-105.)

[5] Gabbay A, Hoshen Y. Scaling-up Disentanglement for Image Translation[C]. *2021 IEEE/CVF International Conference on Computer Vision*, 2022: 6763-6772.

[6] Patashnik O, Wu Z Z, Shechtman E, et al. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery[C]. *2021 IEEE/CVF International Conference on Computer Vision*, 2022: 2065-2074.

[7] Huang S S, Jin X, Jiang Q, et al. A Fully-Automatic Image Colorization Scheme Using Improved CycleGAN with Skip Connections[J]. *Multimedia Tools and Applications*, 2021, 80(17): 26465-26492.

[8] Sun P, Lang Y B, Gong J C, et al. Authentication Method for Splicing Manipulation with Inconsistencies in Color Shift[J]. *Journal of Computer-Aided Design & Computer Graphics*, 2017, 29(8): 1408-1415.
(孙鹏, 郎宇博, 巩家昌, 等. 拼接篡改伪造图像的色彩偏移量不一致取证方法[J]. *计算机辅助设计与图形学学报*, 2017, 29(8): 1408-1415.)

[9] Zhang Y, Jin X, Jiang Q, et al. Deepfake Image Detection Method Based on Autoencoder[J]. *Journal of Computer Applications*, 2021, 41(10): 2985-2990.
(张亚, 金鑫, 江倩, 等. 基于自动编码器的深度伪造图像检测方法[J]. *计算机应用*, 2021, 41(10): 2985-2990.)

[10] Maksutov A A, Morozov V O, Lavrenov A A, et al. Methods of Deepfake Detection Based on Machine Learning[C]. *2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering*, 2020: 408-411.

[11] Cozzolino D, Thies J, Rössler A, et al. ForensicTransfer: Weakly-Supervised Domain Adaptation for Forgery Detection[EB/OL]. 2018: arXiv: 1812.02510. https://arxiv.org/abs/1812.02510.

[12] Wang S Y, Wang O, Zhang R, et al. CNN-Generated Images are Surprisingly Easy to Spot… for now[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 8692-8701.

[13] Nataraj L, Mohammed T M, Manjunath B S, et al. Detecting GAN Generated Fake Images Using Co-Occurrence Matrices[J]. *Electronic Imaging*, 2019, 31(5): 532-537.

[14] Matern F, Riess C, Stamminger M. Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations[C]. *2019 IEEE Winter Applications of Computer Vision Workshops*, 2019: 83-92.

[15] Zhao H Q, Wei T Y, Zhou W B, et al. Multi-Attentional Deepfake Detection[C]. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 2185-2194.

[16] Li J M, Xie H T, Li J H, et al. Frequency-Aware Discriminative Feature Learning Supervised by Single-Center Loss for Face Forgery Detection[C]. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 6454-6463.

[17] Zhang X, Karaman S, Chang S F. Detecting and Simulating Artifacts in GAN Fake Images[C]. *2019 IEEE International Workshop on Information Forensics and Security*, 2020: 1-6.

[18] Durall R, Keuper M, Keuper J. Watch your Up-Convolution: CNN Based Generative Deep Neural Networks are Failing to Reproduce Spectral Distributions[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 7887-7896.

[19] Ciftci U A, Demir I, Yin L J. FakeCatcher: Detection of Synthetic Portrait Videos Using Biological Signals[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, (99): 1-1.

[20] Qi H, Guo Q, Juefei-Xu F, et al. DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythms[C]. *The 28th ACM International Conference on Multimedia*, 2020: 4318-4327.

[21] Frank J, Eisenhofer T, Schönherr L, et al. Leveraging Frequency Analysis for Deep Fake Image Recognition[EB/OL]. 2020: arXiv: 2003.08685. https://arxiv.org/abs/2003.08685.

[22] Dzanic T, Shah K, Witherden F. Fourier Spectrum Discrepancies in Deep Network Generated Images[EB/OL]. 2019: arXiv: 1911.06465. https://arxiv.org/abs/1911.06465.

[23] Schönfeld E, Schiele B, Khoreva A. A U-Net Based Discriminator for Generative Adversarial Networks[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 8204-8213.

[24] Miyato T, Kataoka T, Koyama M, et al. Spectral Normalization for Generative Adversarial Networks[EB/OL]. 2018: arXiv: 1802.05957. https://arxiv.org/abs/1802.05957.

[25] Karras T, Laine S, Aila T M. A Style-Based Generator Architecture for Generative Adversarial Networks[C]. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 4396-4405.

[26] Brock A, Donahue J, Simonyan K. Large Scale GAN Training for High Fidelity Natural Image Synthesis[EB/OL]. 2018: arXiv: 1809.11096. https://arxiv.org/abs/1809.11096.

[27] Zhu J Y, Park T, Isola P, et al. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks[C]. *2017 IEEE International Conference on Computer Vision*, 2017: 2242-2251.

[28] Choi Y, Choi M, Kim M, et al. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 8789-8797.

[29] Chen Q F, Koltun V. Photographic Image Synthesis with Cascaded Refinement Networks[C]. *2017 IEEE International Conference on Computer Vision*, 2017: 1520-1529.

[30] Li K, Zhang T H, Malik J. Diverse Image Synthesis from Semantic Layouts via Conditional IMLE[C]. *2019 IEEE/CVF International Conference on Computer Vision*, 2020: 4219-4228.

[31] Karras T, Laine S, Aittala M, et al. Analyzing and Improving the Image Quality of StyleGAN[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 8107-8116.

[32] Park T, Liu M Y, Wang T C, et al. Semantic Image Synthesis with Spatially-Adaptive Normalization[C]. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 2332-2341.

[33] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the Inception Architecture for Computer Vision[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 2818-2826.

[34] He K M, Zhang X Y, Ren S Q, et al. Deep Residual Learning for Image Recognition[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 770-778.

[35] Chollet F. Xception: Deep Learning with Depthwise Separable Convolutions[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 1800-1807.

[36] Afchar D, Nozick V, Yamagishi J, et al. MesoNet: A Compact Facial Video Forgery Detection Network[C]. *2018 IEEE International Workshop on Information Forensics and Security*, 2019: 1-7.

[37] Tan M X, Le Q V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks[EB/OL]. 2019: arXiv: 1905.11946. https://arxiv.org/abs/1905.11946.

黄珊珊 于 2021 年在云南大学软件工程专业获得硕士学位。现在重庆大学软件工程专业攻读博士学位。研究领域为计算机视觉、信息安全、图像处理。研究兴趣包括: 图像合成、伪造图像检测。Email: huangshanshan9633@163.com

金鑫 于 2018 年在云南大学信息与通信工程专业获得博士学位。现任云南大学软件学院副教授、硕士生导师。研究领域为人工神经网络理论与应用、图像处理、遥感信息处理。研究兴趣包括: 图像融合、伪造图像检测。Email: xinxin_jin@163.com

吴楠 于 2020 年在东莞理工学院网络工程专业获得学士学位。现在云南大学网络空间安全专业攻读硕士学位。研究领域为网络安全、计算机视觉。研究兴趣包括图像伪造检测、图像篡改检测、图像处理。Email: deepfaker@mail.ynu.edu.cn

江倩 于 2019 年在云南大学信息与通信工程专业获得博士学位。现任云南大学软件学院副教授、硕士生导师。研究领域为机器学习图像处理、信息安全、生物信息。研究兴趣包括: 伪造图像检测、模糊理论。Email: jiangqian_1221@163.com